iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** I **Month of publication:** January 2023

DOI: https://doi.org/10.22214/ijraset.2023.48865

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

Fake Job Listing Detection Using Machine Learning Approach

Jayesh Fating[1], Jayant Tumdam[2], Aryan Raut[3], Ajinkya Ladke[4], Aditi Shewale[5]
[1, 2, 3, 4, 5]*GH Raisoni College Of Engineering, Nagpur*

Abstract: ***To avoid fraudulent posts for jobs on the internet, an automated tool using machine learning-based classification techniques is proposed in the paper. Different ML Models are used for training the machine for checking fraudulent posts on the web and the results of those models are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts.***

I. INTRODUCTION

There are a lot of job advertisements on the internet, even on reputed job advertising sites, which never seem fake. But after the election, the so-called recruiters start asking for the money and bank details. Many of the candidates fall into their trap and lose a lot of money and their current job sometimes. So, it is better to identify whether a job advertisement posted on the site is real or fake. Identifying it manually is very difficult and almost impossible. We can apply machine learning to train a model for fake job classification.

It can be trained on the previous real and fake job advertisements and it can identify a fake job accurately [7].

1) According to Federal Trade Commission Americans were scammed out of $68 million due to fake business and job opportunities in the first quarter of 2022.
2) There are a lot of job scams because of unemployment there are a lot of websites that connect a recruiter to a suitable candidate, and sometimes fake recruiters post a job posting on the job portal with the the motive to get money this problem occurs with many job portals later people shift to a new portal in search of a real job but the fake recruiters join this portal as well hence in today's world it is important to detect real and fake jobs.
3) According to PwC's Global Economic Crime and Fraud Survey 2022 shows good news: the proportion of organizations experiencing fraud has remained relatively steady since 2018 [3].
4) However, the survey of 1,296 executives across 53 countries and regions found a rising threat from external perpetrators—bad actors that are quickly growing in strength and effectiveness [3].
5) Nearly 70% of organizations experiencing fraud reported that the most disruptive incident came via an external attack or collusion between external and internal sources.

II. PROJECT OVERVIEW

1) As the introduction gives us an idea, that is why we required the platform to normalize the gateway for our better understanding of detecting scams, hence we have created a model based on NLP and Machine Learning.
2) We have taken the datasets from Kaggle. It a subsidiary of Google LLC is an online community of data scientists and machine learning practitioners.
3) The model contains a feature that describes the posting as either real or not.

Fake job postings are a very small fraction of this dataset, which is as expected because we don't want too much dataset of fake postings.

There are five steps in our model which will follow the NLP pipeline process:-

a) Problem definition
b) Data collection
c) Data cleaning and exploring Pre-processing
d) Modelling
e) Evaluation

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

III. RELATED WORK

Nowadays due to rise in the internet activities, scammers also post fake offers on various social sites like email, Instagram, Facebook, reviews spam, etc.

A. Review Spam Detection

People often post regarding the services they buy on a platform of interaction, but spammers can manipulate the information and attract other new users by showing their interest in them. This can be extracted from features of NLP and ML where set dictionary features are available to detect spam.

B. E-mail Spam

Useless and unnecessary mail received from unknown or known sources often arrived in the user's mail as email spam. And the result of this is unnecessary feed-up of space of the user. To resolve this issue various email providers like Gmail, outlook provides spam filtering services that use neural network method.

C. Fake News Detection

Any interesting news seen on social media fascinates users towards the news. And mostly people rely on the news seen on social media platforms they never overcheck it, which often results in scams.

1) *To safeguard oneself from these scams, PIB provided three tips:* Avoid clicking on unverified links, regardless of how tempting they might appear. Be cautious before conducting any monetary transaction with strangers

The best option is to report and block such numbers.

IV. PROPOSED METHODOLOGY

The target of this study is to detect whether a job post is fraudulent or not. Identifying and eliminating these fake job advertisements will help the job-seekers to concentrate on legitimate job posts only. In this context, a dataset from Kaggle is employed that provides information regarding a job that may or may not be suspicious.

The following steps are taken for text processing:

1) *Tokenization:* The textual data is split into smaller units. In this case, the data is split into words.
2) *To Lower:* The split words are converted to lowercase
3) *Stopword Removal:* Stopwords are words that do not add much meaning to sentences. For example the, a, an, he, have, etc. These words are removed.
4) *Lemmatization:* The process of lemmatization groups in which inflected forms of words are used together.

The Dataset has the following schema which is as below:-

```
];  job_id                   0
    title                    0
    location               346
    department           11547
    salary_range         15012
    company_profile       3308
    description              1
    requirements          2695
    benefits              7210
    telecommuting            0
    has_company_logo         0
    has_questions            0
    employment_type       3471
    required_experience   7050
    required_education    8105
    industry              4903
    function              6455
    fraudulent               0
    dtype: int64
```


International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

V. MODEL CALCULATION

The models will be evaluated based on two metrics:

1) Accuracy: This metric is defined by this formula -

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{False Negative} + \text{True Negative}}$$

As the formula suggests, this metric produces a ratio of all correctly categorized data points to all data points. This is particularly useful since we are trying to identify both real and fake jobs unlike a scenario where only one category is important. There is however one drawback to this metric. Machine learning algorithms tend to favor dominant classes. Since our classes are highly unbalanced a high accuracy would only be a representative of how well our model is categorizing the negative class (real jobs).

2) F1-Score: The F1 score is a measure of a model's accuracy on a dataset. The formula for this metric is –

$$F_1 = \frac{\text{True Positive}}{\text{True Positive} + 1/2(\text{False Positive} + \text{False Negative})}$$

F1-score is used because in this scenario both false negatives and false positives are crucial. This model needs to identify both categories with the highest possible score since both havehigh costs associated with them.

VI. RESULTS

A. Model Evaluation and Validation

The final model used for this analysis is – SGD. This is based on the results of the metrics as compared to the baseline model. The outcome of the baseline model and SGD are presented in the table below:

Model	Accuracy	F1-score
Naïve Bayes (baseline model)	0.971	0.743
SGD	0.974	0.79

Based on these metrics, SGD has a slightly better performance than the baseline model. This is how the final model is chosen to be SGD.

VII. CONCLUSION

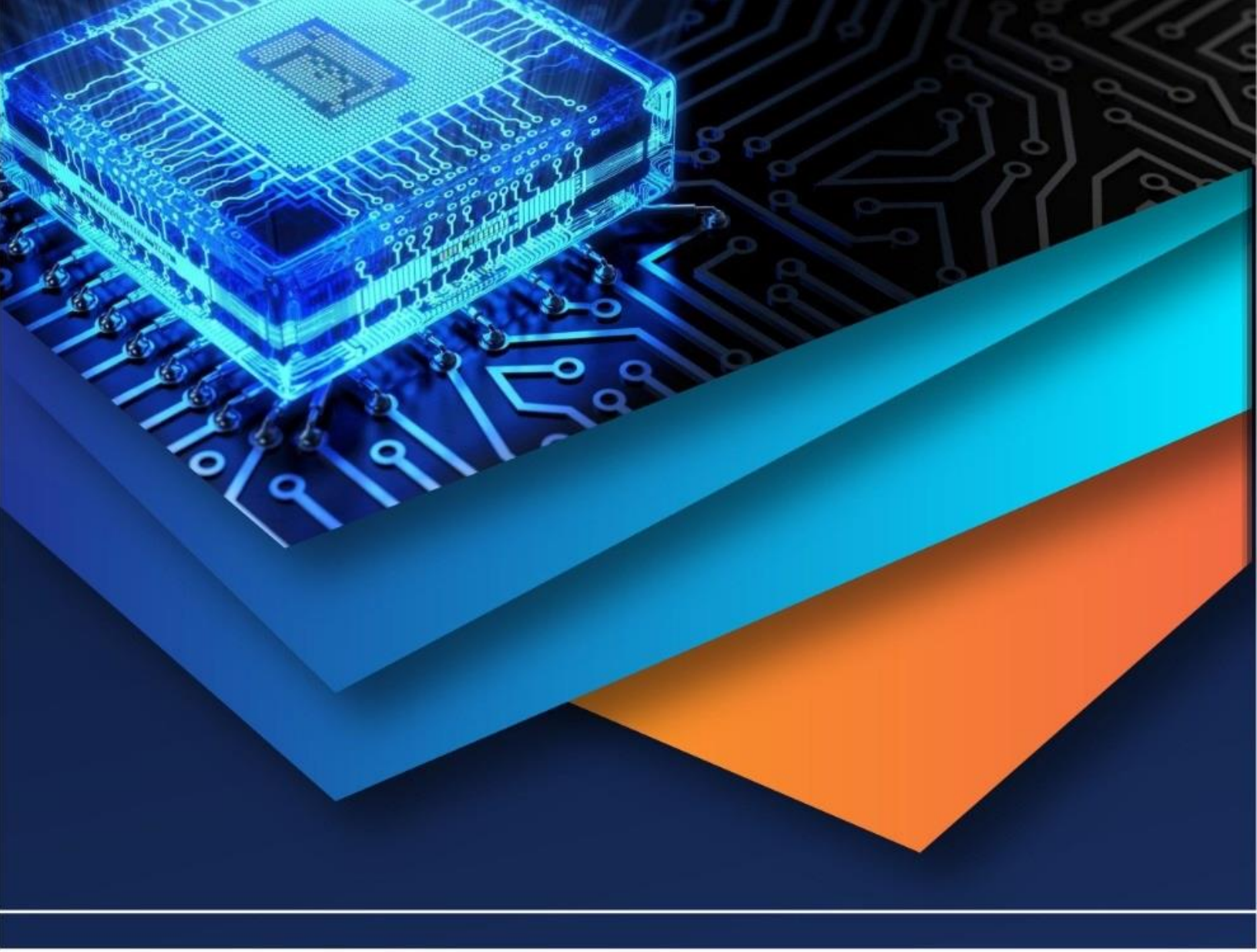
A. Free-Form Visualization

A confusion matrix can be used to evaluate the quality of the project. The project aims to identify real and fake jobs.

The confusion matrix above displays the following values – categorized label, number of data points categorized under the label, and percentage of data represented in each category. The test set has a total of 3265 real jobs and 231 fake jobs. Based on the confusion matrix it is evident that the model identifies real jobs 99.01% of the time. However, fraudulent jobs are identified only 73.5% of the time. Only 2% of the time has the model not identified the class correctly. This shortcoming has been discussed earlier as well as Machine Learning algorithms tend to prefer the dominant classes.

REFERENCES

[1] Fake Job Recruitment Detection Using Machine Learning Approach- Shawni Dutta and Prof. Samir Kumar Bandyopadhyay International Journal of Engineering Trends and Technology (IJETT) – Volume 68 Issue 4- April 2020 .
[2] Fake Job Posting Prediction using machine learning - Anshupriya Srivastava- Github
[3] PwC's Global Economic Crime and Fraud Survey- https://www.pwc.com/gx/en/services/forensics/economic-crime-survey.html
[4] Kaggle Datasets- - https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting- prediction
[5] D. E. Walters, ―Bayes's Theorem and the Analysis of Binomial Random Variables,‖ Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710
[6] Edureka fake job lisiting detection using ML And NLP.
[7] Analytics India Mag- https://analyticsindiamag.com/classifying-fake-and-real-job- advertisements-using-machine-learning/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY